Case 3:23-cv-03417-VC Document 504-4 Filed 03/28/25 Page 1 of 6

EXHIBIT C

<u>L1189420PRV</u>: Meta Large Language Model (LLM): Approving Existing Datasets Public Launch

LaMa ID: L1189420PRV

Summary:

The GenAl team is seeking approval to leverage 3P pre-training dataset(s) for the purpose of further training the Llama model/LLM powered chatbot for public launch.

Note: Public Launch of Meta "Chatbot" on WhatsApp will require a separate PXFN review.

Datasets/Training Data:

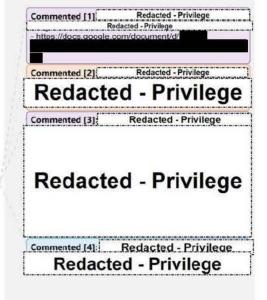
- Stack Exchange (Top 28)
- Bocks3
- Gutenberg
- Arxiv
- · Github open licenses (MIT/Apache)
- C4 (Common crawl)
- CCNet (Common Crawl)
- CC-stories (Common crawl)
- The Stack
- Wikipedia
- Math

Model Training & Data flow:

After downloading the data, we apply deduplication within each dataset, and filter out high-risk IP and PII websites, specifically the list here: a/c priv - Third-Party GenAl LLM Training Dataset PII & Scraping Principles See more info here: LLM Safety: PII Mitigations - [A/C Priv] Proposed IP Mitigations for LLM Chatbot

Next, the GenAl ML team will extract tokens, features, from the text data (i.e., words, clips, sentences, paragraphs), which will be fed to the model to learn patterns of speech and writing.

Training data will have a retention period of 3 years, stored in Hive with access restricted via ACL to the LLM Teams..



Note: The team has added purchased data (human annotations) into the training mix.

Scope:

This privacy decision approves the use of the aforementioned 3P datasets for the purpose of further training the Llama model/LLM powered chatbot for public launch.

Category	Mitigation
Purpose Limitation	3P Datasets must only be used for the purpose of further training the Llama model/LLM powered chatbot. Team must delete highest risk urls w/in common crawl datasets. Data has been deduplicated, which has been shown in academic literature to reduce memorization
Data Retention	 *update* Training data may be retained until training is complete, or until a user request their data is removed via the objection form, or 3 years whichever is earlies datasets must be deleted from our storage systems and not used in future model training if complaints to remove or correct certain data have been received within 60 days.

Internal Unauthorized Access	 Training data must be stored securely and restricted to the LLM team by access control list.
User Access & Management Deletion	
Other	

Document Comments
Total Comments: 4
Author: Ahuva Goldstand
Date: 5/11/2023 9:08:00 AM
Range: Redacted - Privilege
- https://docs.google.com/document/d/
Scope: After downloading the data, we apply deduplication within each dataset, and filter out hig
risk IP and PII websites, specifically the list here: a/c priv - Third-Party GenAI LLM Training
Dataset PII & Scraping Principles See more info here: LLM Safety: PII Mitigations
Author: Dustin Holland
Author: Dustin Holland Date: 5/11/2023 9:06:00 PM Range: Redacted - Privilege
Range : Redacted - Privilege Redacted - Privilege
Scope: After downloading the data, we apply deduplication within each dataset, and filter out hig
risk IP and PII websites, specifically the list here: a/c priv - Third-Party GenAI LLM Training Dataset PII & Scraping Principles See more info here: LLM Safety: PII Mitigations
bacaset PII & Scraping Principles See more into here: him Safety: PII Micigations
Author: Ahuva Goldstand
Date: 5/15/2023 11:40:00 AM
Range: Redacted - Privilege
DAMAATA Drivilada
Redacted - Privilege
110000000 111111090
Scope: After downloading the data, we apply deduplication within each dataset, and filter out hig
risk IP and PII websites, specifically the list here: a/c priv - Third-Party GenAI LLM Training

Dataset PII & Scraping Principles See more info here: LLM Safety: PII Mitigations

Author: Melanie Kambadur Date: 5/15/2023 7:33:00 PM

Redacted - Privilege Range:

Redacted - Privilege

Scope: After downloading the data, we apply deduplication within each dataset, and filter out high-risk IP and PII websites, specifically the list here: a/c priv - Third-Party GenAI LLM Training Dataset PII & Scraping Principles See more info here: LLM Safety: PII Mitigations